doi: 10.1111/j.1365-313X.2006.02829.x

Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton

S. Samuel Yang[1,†], Foo Cheung[2,†], Jinsuk J. Lee[3], Misook Ha[3], Ning E. Wei[4], Sing-Hoi Sze[4], David M. Stelly[1], Peggy Thaxton[5], Barbara Triplett[6], Christopher D. Town[2] and Z. Jeffrey Chen[1,3,*]

[1]*Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843, USA,*
[2]*The Institute for Genomic Research, Rockville, MD 20850, USA,*
[3]*Section of Molecular Cell and Developmental Biology, The University of Texas, Austin, TX 78712, USA,*
[4]*Department of Computer Science, Texas A&M University, College Station, TX 77843, USA,*
[5]*Delta Research and Extension Center, Mississippi State University, Stoneville, MS 38776, USA, and*
[6]*USDA-ARS Southern Regional Research Center, New Orleans, LA 70179, USA*

Received 22 February 2006; revised 15 May 2006; accepted 25 May 2006.

[*]For correspondence (fax +1 512 232 3432; e-mail zjchen@mail.utexas.edu).
[†]These two authors contributed equally to this work.

Summary

Gene expression during the early stages of fiber cell development and in allopolyploid crops is poorly understood. Here we report computational and expression analyses of 32 789 high-quality ESTs derived from *Gossypium hirsutum* L. Texas Marker-1 (TM-1) immature ovules (GH_TMO). The ESTs were assembled into 8540 unique sequences including 4036 tentative consensus sequences (TCs) and 4504 singletons, representing approximately 15% of the unique sequences in the cotton EST collection. Compared with approximately 178 000 existing ESTs derived from elongating fibers and non-fiber tissues, GH_TMO ESTs showed a significant increase in the percentage of genes encoding putative transcription factors such as MYB and WRKY and genes encoding predicted proteins involved in auxin, brassinosteroid (BR), gibberellic acid (GA), abscisic acid (ABA) and ethylene signaling pathways. Cotton homologs related to *MIXTA*, *MYB5*, *GL2* and eight genes in the auxin, BR, GA and ethylene pathways were induced during fiber cell initiation but repressed in the naked seed mutant (*N1N1*) that is impaired in fiber formation. The data agree with the known roles of MYB and WRKY transcription factors in Arabidopsis leaf trichome development and the well-documented phytohormonal effects on fiber cell development in immature cotton ovules cultured *in vitro*. Moreover, the phytohormonal pathway-related genes were induced prior to the activation of *MYB*-like genes, suggesting an important role of phytohormones in cell fate determination. Significantly, AA sub-genome ESTs of all functional classifications including cell-cycle control and transcription factor activity were selectively enriched in *G. hirsutum* L., an allotetraploid derived from polyploidization between AA and DD genome species, a result consistent with the production of long lint fibers in AA genome species. These results suggest general roles for genome-specific, phytohormonal and transcriptional gene regulation during the early stages of fiber cell development in cotton allopolyploids.

Keywords: cotton, expressed sequence tags, fiber, gene expression, phytohormone, polyploidy.

Introduction

The most extensively cultivated cotton species are allotetraploid *Gossypium hirsutum* L. (upland or American cotton) and *G. barbadense* L. ('Egyptian' cotton). Both allotetraploids originated in the New World from inter-specific hybridization between species closely related to *G. herbaceum* L. (A_1) or *G. arboreum* L. (A_2) and an American diploid, *G. raimondii* L. (D_5) or *G. gossypioides* (Ulbrich) Standley (D_6; Beasley, 1940). This polyploidization event is estimated to have occurred 1–2 million years ago (Wendel and Cronn, 2003; Wendel *et al.*, 1995), and gave rise to a disomic allopolyploid consisting of five extant allotetraploid species (Percival *et al.*, 1999). The AA progenitor species produce both lint (long) fibers that may be spun into yarn and shorter fibers called fuzz. Lint fibers usually initiate on the day of anthesis, and fuzz fibers develop at a later stage. In contrast, the DD genome progenitor species produce a few lint fibers that are initiated pre-anthesis, but these are much shorter in length than the lint fibers of the AA genome progenitor (Applequist *et al.*, 2001). Compared with the AA and DD genome progenitors, the fiber traits in the allotetraploids are dramatically enhanced, suggesting that inter-genomic interactions induce tissue-specific expression of homoeologous genes in cotton allotetraploids (Adams *et al.*, 2003).

Cotton fibers are seed trichomes. During cotton fiber development, protodermal cells of ovules undergo several distinctive but overlapping steps, including fiber initiation, elongation, secondary cell wall biosynthesis, and maturation, leading to mature fibers (Basra and Malik, 1984; Kim and Triplett, 2001; Tiwari and Wilkins, 1995; Wilkins and Jernstedt, 1999). In *G. hirsutum*, lint fibers develop prior to or on the day of anthesis, and the process is quasi-synchronized in each developing ovule and among ovules within each ovary (boll). Fuzz fiber development usually occurs in a later stage but this varies among genotypes.

Transcription factors such as WD40 proteins (*TTG1*), MYB (*GL1* or *WER*), and basic helix-loop-helix proteins (*GL3* or *EGL3*) play a role in determining epidermal trichome cell patterning in Arabidopsis leaves (Glover, 2000; Hülskamp, 2004; Ramsay and Glover, 2005). This complex is thought to activate a homeodomain leucine zipper protein (*GL2*) and a small family of single-repeat MYB proteins without transcription activation domains (*TRY*, *CPC* and *ETC1*). Interestingly, *GL2* is an activator of downstream trichome-specific differentiation genes, whereas *TRY* (*CPC* or *ETC1*) is a negative regulator that represses trichome differentiation by competing with the MYB factors for binding of the initiation complex.

Similar genes and pathways may be involved during seed trichome development in cotton, although cotton fibers are unicellular and never branch. *GhMYB109*, a putative ortholog of *AtMYBGL1*, is specifically expressed in cotton fiber cell initials and elongating fibers (Suo *et al.*, 2003). *GaMYB2*, another R2R3 MYB transcription factor related to *AtMYBGL1*, complements the *Arabidopsis glabrous1* (*gl1*) mutant. Moreover, ectopic expression of *GaMYB2* induces a single trichome from the epidermis of Arabidopsis seed (Wang *et al.*, 2004), and two cotton genes containing WD40 domains complement the Arabidopsis *ttg1* mutant (Humphries *et al.*, 2005). *GhMYB25*, a homolog of *AmMIXTA*/*AmMYBML1* that controls conical cell and trichome differentiation in *Antirrhinum majus* petals (Noda *et al.*, 1994; Perez-Rodriguez *et al.*, 2005), is predominately expressed in ovules and in fiber cell initials (Wu *et al.*, 2006).

A recent study using microarray and quantitative gene expression analyses has indicated that ethylene is involved in fiber cell elongation (Shi *et al.*, 2006). Moreover, BR promotes fiber cell development on cultured cotton ovules (Sun *et al.*, 2005) in a manner reminiscent of the well-established requirement for plant hormones (GA and auxin; Beasley and Ting, 1974). Collectively these data suggest critical roles for phytohormones in fiber cell development.

Obviously, cotton fiber cell initiation is a complex biological process that requires orchestrated changes in gene expression in developmental and physiological pathways (Arpat *et al.*, 2004; Ji *et al.*, 2003; Kim and Triplett, 2001; Lee *et al.*, 2006; Li *et al.*, 2002; Wilkins and Arpat, 2005). Many candidate genes that are expressed in fiber cells have been cloned and characterized (Delmer *et al.*, 1995; John and Keller, 1995; Kim and Triplett, 2004; Orford and Timmis, 1997; Reinhart *et al.*, 1996; Suo *et al.*, 2003). For example, the expression of some genes is associated with the fiber elongation stage of development (John and Crow, 1992; Kim and Triplett, 2004; Ma *et al.*, 1997; Orford and Timmis, 1997; Smart *et al.*, 1998; Suo *et al.*, 2003), whereas others are preferentially expressed during secondary cell wall thickening (Haigler *et al.*, 2005; John and Keller, 1995; Reinhart *et al.*, 1996; Wilkins and Jernstedt, 1999), or constitutively expressed throughout fiber development (Whittaker and Triplett, 1999).

The molecular events during fiber cell initiation are poorly understood. As of April 2006, the cotton EST collection in the public database (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=cotton) contained about 211 028 ESTs largely derived from *G. arboreum* L. (Arpat *et al.*, 2004), *G. hirsutum* L. (Haigler *et al.*, 2005; Li *et al.*, 2002; http://www.biology.bnl.gov/plantbio/burr.html) and *G. raimondii* Ulbrich (Udall *et al.*, 2006; http://genome.arizona.edu/genome/cotton.html). The majority of fiber ESTs are derived from fibers in the early elongation stage (Arpat *et al.*, 2004) or from the secondary wall synthesis stage of fiber development (Haigler *et al.*, 2005). Therefore, new cotton EST sequences derived from tissues in the earlier stages of development from 3 days pre-anthesis (−3 DPA) to 3 days post-anthesis (+3 DPA)[‡] are essential for uncovering additional genes involved in the complex biological networks leading to fiber cell differentiation. Here we report characterization of 32 798 ESTs derived from immature and fiber-bearing ovules in comparison with approximately 178 000 other cotton ESTs in the database. The data indicate that (1) a large number of ESTs are differentially represented in fibers and non-fiber tissues, (2) G. hirsutum L. Texas Marker-1 (TM-1) immature ovules (GH_TMO) ESTs are highly

[‡] The stages of ovule development are referenced relative to anthesis, also called 0 days post-anthesis. The number of days pre-anthesis is designated with a negative value and the number of days post-anthesis is designated as a positive value.

enriched with genes encoding putative transcription factors and phytohormonal regulators, and (3) many AA subgenome ESTs are selectively enriched in *G. hirsutum* L. TM-1. A subset of genes encoding putative MYB transcription factors and auxin, BA, GA and ethylene regulators is induced during early stages of fiber cell development in TM-1 but repressed in the *N1N1* mutant that produces very few lint fibers and no fuzz fibers. These results suggest important roles for transcription factors, phytohormonal regulators and genome-specific gene regulation in the early stages of fiber cell development.

Results

The GH_TMO full-length cDNA library contained 4.2×10^6 colony-forming units with approximately 99% colony recovery and >60% full-length cDNA inserts, with an average insert size of 1.53 kb. A total of 32 789 high-quality EST sequences were obtained after removal of vector, poly-A and contaminating microbial sequences. The average EST length was 763 bp, and approximately 78% were >700 bp. The average length of GH_TMO ESTs was approximately 120 bp longer than that of other cotton ESTs in the database.

EST sequence assembly, annotation and cluster analysis

The ESTs were assembled into 8540 unique sequences, consisting of 4036 tentative consensus sequences (TCs) and 4504 singletons. The average length was 881 bp for all unique sequences, 1050 bp for TCs and 730 bp for singletons. The Cotton Gene Index version 6 (CGI6, http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=cotton) as of April 2006 contained 40 348 unique sequences. We revised CGI6 using all EST (211 397) sequences downloaded from The National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/), including those generated in this study. The resulting CGI7 contained 55 673 unique sequences, of which 21 405 were TCs and 34 268 singletons. The average length was 834 bp for unique sequences, 1077 bp for TCs and 590 bp for singletons.

Approximately 15% of the total unique sequences in CGI7 were derived exclusively from the GH_TMO library, of which 2686 (approximately 4.8%), including 654 TCs and 2032 singletons, were GH_TMO-specific transcripts that were enriched in ovules during the earliest stage of fiber development (from −3 to +3 DPA) compared with other ESTs (Table S1).

The putative functions for all unique sequences were assigned using BLAST searches against the non-redundant (NR) protein database. About 26% of the sequences in CGI7 were putative cotton-specific sequences (no hits). Table 1 indicates the 30 most abundantly expressed TCs, encoding predicted proteins such as protodermal factors (*TC1* and *TC2*), peroxidase (*TC4*), mitochondrial carrier protein (*TC23*), α-expansin (*TC12*) and E6-fiber protein (*TC47*). Many of these abundant transcripts were found only in the GH_TMO library.

Table 1 The 30 most abundant transcripts in GH_TMO

TC number[a]	No. ESTs[b]	Putative function
TC1	502	Protodermal factor 1
TC4	303	Peroxidase precursor
TC2	268	Protodermal factor 1
TC23	172	Mitochondrial carrier protein family
TC12	158	α-expansin precursor
TC2034	151	Pentameric polyubiquitin
TC6	147	α-tubulin
TC10	130	Adenosylhomocysteinase
TC24	129	Cytochrome P450-like TBP thylakoid-binding protein
TC2026	129	*S*-adenosylmethionine synthetase 2
TC2050	123	Flavonoid 3′,5′-hydroxylase
TC2047	119	Expressed protein (At5g12010)
TC2041	115	Chalcone synthase 1
TC2044	111	Fructose-bisphosphate aldolase
TC2052	110	Flavanone 3-hydroxylase
TC25	109	Heavy-metal-associated domain-containing protein
TC34	108	Phi-1 protein
TC11	94	Adenosylhomocysteinase
TC38	90	Expressed protein (At1g09750)
TC7	85	Tubulin α4 chain
TC2035	83	Polyubiquitin
TC2055	82	Tuber-specific and sucrose-responsive element binding factor
TC2042	82	Chalcone synthase 1
TC2039	80	Histone H1
TC2053	78	60S acidic ribosomal protein P0
TC42	74	Inositol-3-phosphate synthase
TC2070	72	Cytochrome P450
TC18	72	Elongation factor 1-α
TC2071	70	60S ribosomal protein L4
TC47	70	E6

[a] TC numbers are from the GH_TMO unique sequence set (Table S2).
[b] The number of ESTs present in each TC.

To determine the functional distributions of ESTs, we analyzed the percentage of gene ontology (GO) molecular function classes for the GH_TMO ESTs, CGI6 and CGI7 gene indices, and the Arabidopsis proteome database. Compared with the number of annotated genes in each GO functional class in the Arabidopsis genome and CG16, GH_TMO ESTs showed a significant increase ($P \leq 0.01$, χ^2-test) in the percentage of predicted genes in the classes of transcription factor activity, DNA or RNA binding, nucleic acid binding, and nucleotide binding (Figure 1).

Enrichment of transcription factors in the GH_TMO ESTs

To determine the relative abundance of putative transcription factors in the cotton EST collection, we compared 1827 protein sequences consisting of 56 transcription factor families from the Arabidopsis transcription factor database

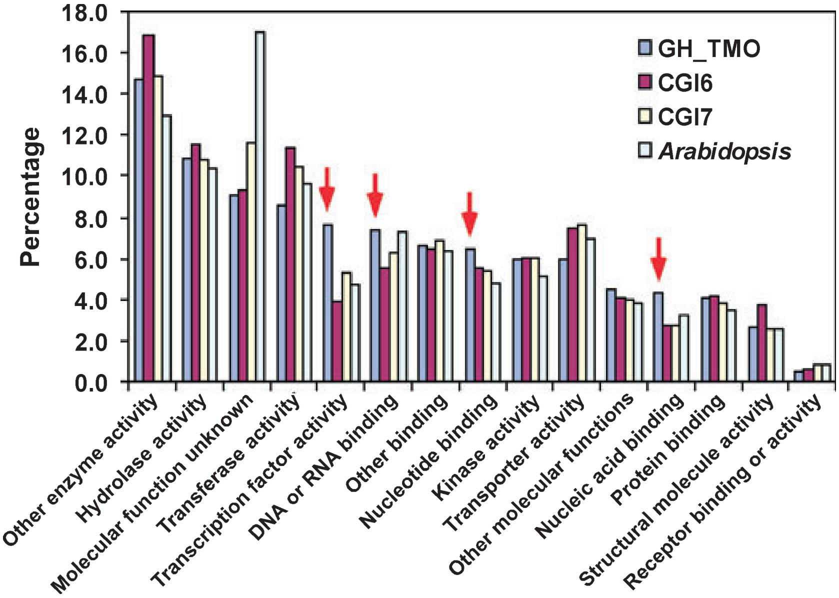


Figure 1. Gene ontology (GO) molecular functional classifications of GH_TMO, CGI6, CGI7 and the Arabidopsis proteome.
The GO functional class was assigned to cotton unique sequences using Arabidopsis GO SLIM (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.20050723.txt). Arrows indicate the GO functional classes that contained ESTs which were significantly over-represented in the GH_TMO library using χ^2 tests.

(http://datf.cbi.pku.edu.cn/index.php) with the cotton EST sequences using TBLASTN ($E \leq -10$). Notably, the frequency of putative transcription factors in the GH_TMO library (approximately 10%) was significantly higher than that in CGI6 (approximately 4.7%), CGI7 (approximately 5.0 %) and the Arabidopsis proteome (approximately 6.3%; $P \leq 0.01$, χ^2 test; Figure 2), a result consistent with the GO functional classification data described above.

Among the putative transcription factor sequences identified in CGI7 (Table S2), GH_TMO ESTs contained a total of 251 unique sequences, including 94 TCs and 157 singletons. Almost every putative transcription factor family (with the exception of MADS) was over-represented in the GH_TMO library (Figure 2). Many ESTs encoding putative transcription factors, including MYB, WRKY, AP2/EREBP, C2H2 and bHLH families, were exclusively present in the GH_TMO library.

CGI7 contained a total of 242 putative MYB-coding sequences, 21 of which (approximately 8.7%) are specific to the GH_TMO library (Table S2). The cotton MYB-like genes are distributed among various clades and subgroups in the neighbor-joining phylogenetic tree, and some are located in the clades specific to cotton (Figure 3). Several putative cotton MYB orthologs matched *PhMYB1* (Z13996) and *AmMIXTA* (X79108), which play a role in leaf trichome development (Noda *et al.*, 1994; Perez-Rodriguez *et al.*,

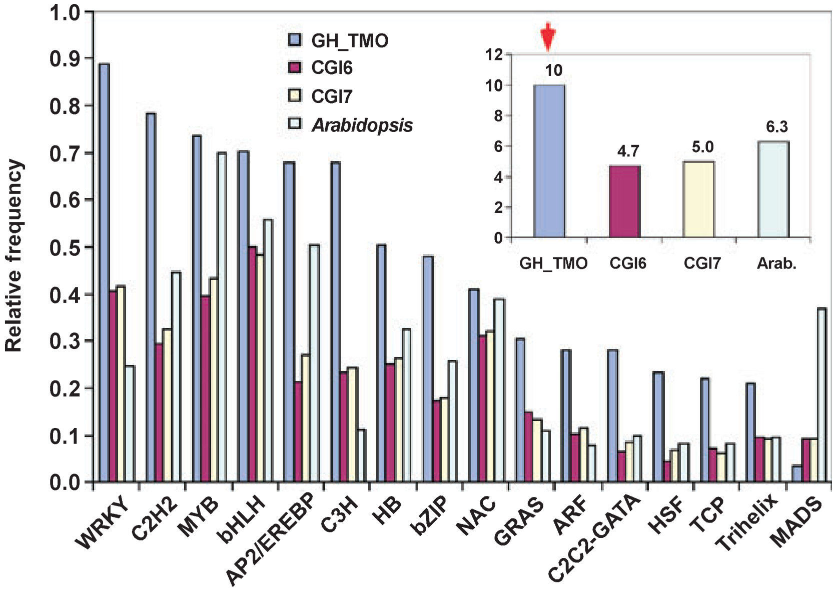


Figure 2. Relative frequencies of 16 putative transcription factor families present in GH_TMO, CGI6, CGI7 and the Arabidopsis proteome.
The inset box shows the percentage of all putative transcription factors present in GH_TMO, CGI6, CGI7 and the Arabidopsis (Arab.) proteome, respectively. The arrow indicates that 10% of the putative transcription factors are present in the GH_TMO library.

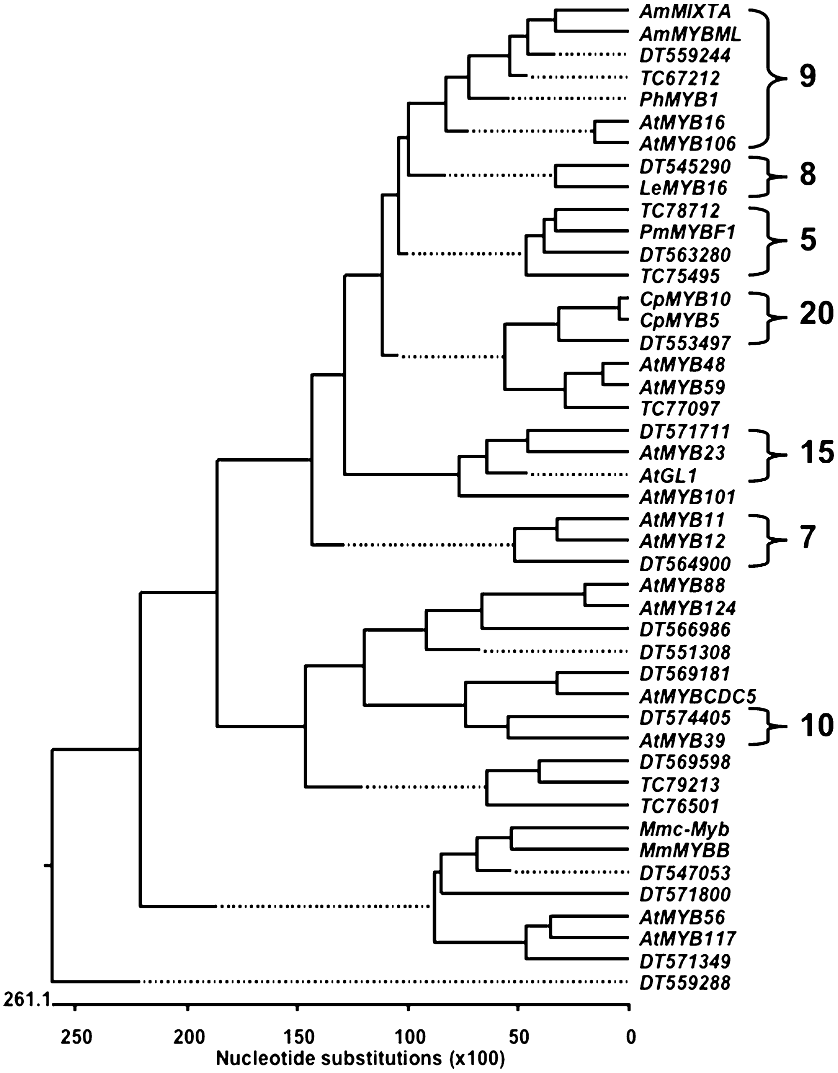
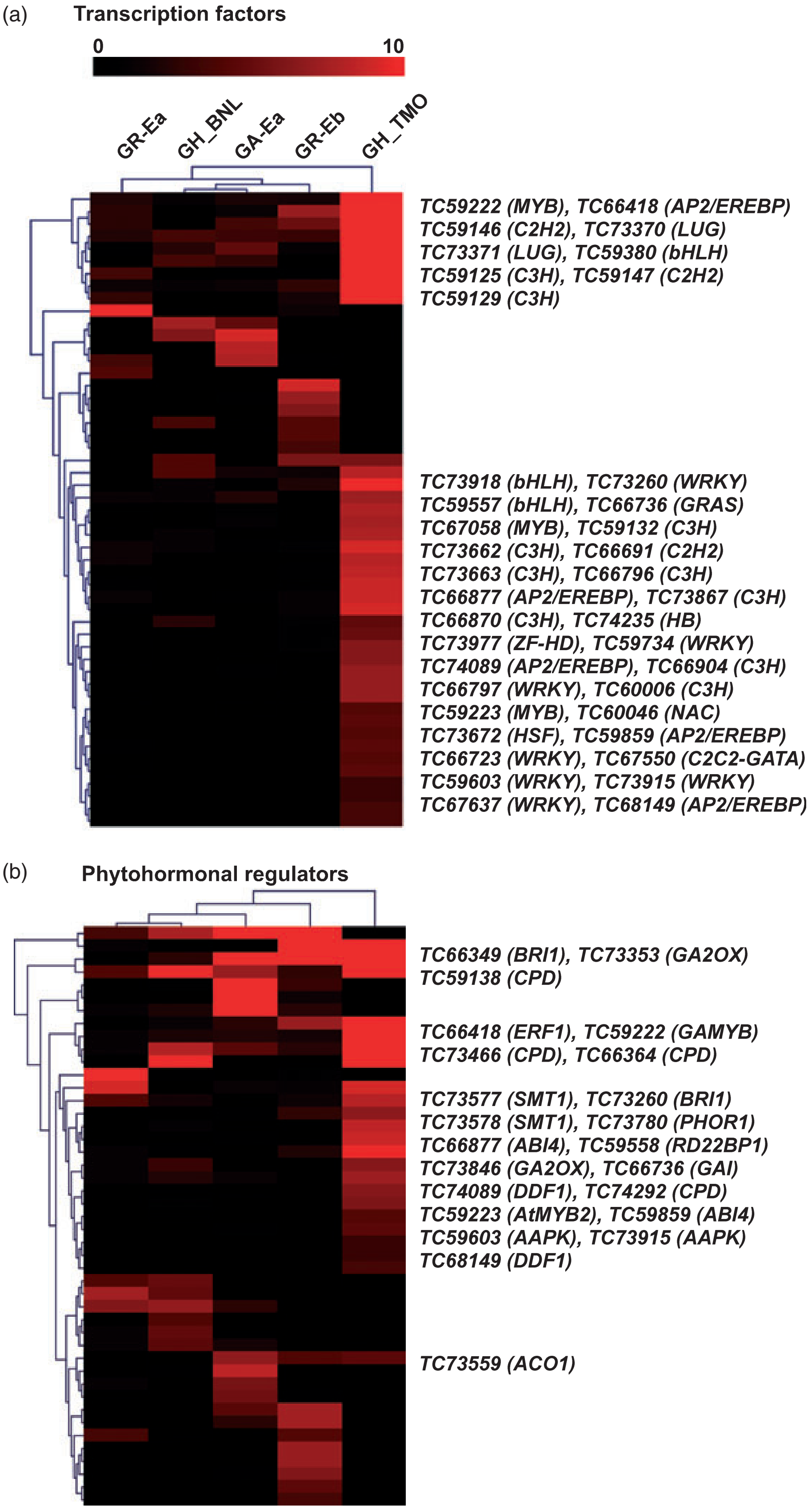


Figure 3. A phylogenetic tree indicating the relationships of 21 GH_TMO-unique sequences encoding putative MYB factors and 24 genes encoding putative MYB factors in other plant species.
TCs and singletons (DTs) were based on CGI7 (Table S3). Sequences other than those for the Arabidopsis MYB genes are shown with the gene names. Subgroups are designated according to the previously described method (Kranz *et al.*, 1998).

2005). A total of 197 unique sequences encoding putative WRKY factors were identified in CGI7, 32 of which (approximately 16%) were present only in the GH_TMO library.

To determine the relative enrichment of ESTs in different tissues, we analyzed EST data generated in five EST libraries using *R* statistics (Stekel *et al.*, 2000) with a *P* value of 0.001 (Table S1). The libraries were derived from GH_TMO (*G. hirsutum* TM-1 ovules), GH_BNL (*G. hirsutum* s-day cotton fibers), GA_Ea (*G. arboreum* developing fibers; 7–10 DPA), GR_Ea (*G. raimondii* whole seedlings with the first true leaves), and GR_Eb (*G. raimondii* bolls; –3 DPA flower buds to +3 DPA bolls). A total of 648 genes showed differential representation in the five libraries. Using hierarchical cluster analysis (Eisen *et al.*, 1998), we found that many differentially enriched genes were library-specific (Figure 4), suggesting a relative enrichment of these genes in different tissues. Notably, ESTs related to several clusters of transcription factors were more abundant in the GH_TMO library, but not in other libraries. In addition to MYBs, ESTs encoding many other putative transcription factors, including MADS, C2H2, C3H, bHLH and WRKY, were highly represented in the GH_TMO library (Figure 4a).

Using quantitative RT-PCR analysis, we confirmed the transcript abundance of selected ESTs encoding transcription factors in ovules at early stages of fiber development (–3 to +3 DPA; Figure 5). *DT559244* and *TC67212* are putative cotton orthologs of *AmMIXTA*/*AmMYBML1*. The expression of these two genes was rapidly induced at 0 to +3 DPA and declined at 5 DPA (Figure 5a,b). *DT556361*, a putative cotton ortholog of *AtGL2*, was expressed at high levels in fiber-bearing ovules but at significantly reduced levels in non-fiber tissues (Figure 5c). *DT553497*, a putative cotton ortholog of *CpMYB5*, was highly expressed at –3 and 0 DPA but dramatically downregulated after +3 DPA (Figure 5d). Interestingly, the expression of all four genes was significantly downregulated in the *N1N1* mutant compared


(a)
Transcription factors
0
10
GR-Ea
GH_BNL
GA-Ea
GR-Eb
GH_TMO
TC59222 (MYB), TC66418 (AP2/EREBP)
TC59146 (C2H2), TC73370 (LUG)
TC73371 (LUG), TC59380 (bHLH)
TC59125 (C3H), TC59147 (C2H2)
TC59129 (C3H)
TC73918 (bHLH), TC73260 (WRKY)
TC59557 (bHLH), TC66736 (GRAS)
TC67058 (MYB), TC59132 (C3H)
TC73662 (C3H), TC66691 (C2H2)
TC73663 (C3H), TC66796 (C3H)
TC66877 (AP2/EREBP), TC73867 (C3H)
TC66870 (C3H), TC74235 (HB)
TC73977 (ZF-HD), TC59734 (WRKY)
TC74089 (AP2/EREBP), TC66904 (C3H)
TC66797 (WRKY), TC60006 (C3H)
TC59223 (MYB), TC60046 (NAC)
TC73672 (HSF), TC59859 (AP2/EREBP)
TC66723 (WRKY), TC67550 (C2C2-GATA)
TC59603 (WRKY), TC73915 (WRKY)
TC67637 (WRKY), TC68149 (AP2/EREBP)
(b)
Phytohormonal regulators
TC66349 (BRI1), TC73353 (GA2OX)
TC59138 (CPD)
TC66418 (ERF1), TC59222 (GAMYB)
TC73466 (CPD), TC66364 (CPD)
TC73577 (SMT1), TC73260 (BRI1)
TC73578 (SMT1), TC73780 (PHOR1)
TC66877 (ABI4), TC59558 (RD22BP1)
TC73846 (GA2OX), TC66736 (GAI)
TC74089 (DDF1), TC74292 (CPD)
TC59223 (AtMYB2), TC59859 (ABI4)
TC59603 (AAPK), TC73915 (AAPK)
TC68149 (DDF1)
TC73559 (ACO1)

with TM-1. The *N1N1* mutation leads to a naked seed phenotype due to delayed fiber cell initiation, decreased number of fiber cell initials, a reduced rate of fiber elongation, and the absence of long lint fibers (Endrizzi *et al.*, 1984; Lee *et al.*, 2006).

Enrichment of phytohormonal regulators in the GH_TMO ESTs

To test the role of hormonal signaling in early stages of ovule and fiber development, we analyzed cotton EST sequences using TBLASTN ($E \leq -10$) against the amino acid sequences of Arabidopsis gene families involved in phytohormone biosynthesis and signal transduction pathways. We identified 230 putative ABA-, BR-, GA-, ethylene- and auxin-related sequences in the GH_TMO library (Figure 4b and Table S3). Moreover, the hierarchical cluster analysis indicated that several subsets of putative phytohormonal pathway genes were differentially enriched in five cotton EST libraries, many of which accumulated only in ovules from −3 to +3 DPA (Figure 4b). The GH_TMO library was enriched with the ESTs encoding putative cotton BR-associated genes in the biosynthetic (*SMT1*, *SMT2s* and *BR6OX*) and signaling (*BRI1s*, *BAK1s*, *BES1s* and *CPD*) pathways (Figure 4b and Table S3). Two putative cotton *BES1* orthologs, a positive regulator in the downstream BR signaling pathway (Yin *et al.*, 2002), were present only in the GH_TMO library, suggesting a role for BR in fiber cell differentiation. Calcium-dependent protein kinases (CDPKs) are implicated in GA and BR signaling in rice (Abo-el-Saad and Wu, 1995; Yang and Komatsu, 2000). Several *CDPK*s (*DT569356*, *TC64112*, *DT572262*, *TC74526* and *TC63968*) were found in the GH_TMO library, whereas others (*TC66333* and *TC66335*) were present in the cDNA libraries derived from elongating fibers (GH_BNL and GA_Ea).

Many GH_TMO ESTs encode putative regulators in the GA biosynthetic (*GA20OX*, *GA2OX*, *POTH1* and *KO*) and signaling (*GAI*, *RGL2*, *RGL1*, *DDF1*, *PHOR1*, *RSG*, *PKL*, *GL1*, *GAMYB*, *AGAMOUS* and *LUE1*) pathways (Figure 4b and Table S3). For example, putative cotton homologs of *POTH1* (Hedden and Kamiya, 1997) and the GA2-oxidase gene (Rosin *et al.*, 2003) were present only in the GH_TMO library. The GH_TMO library contained seven putative DELLA-like genes (four *RGL2*s, two *GAI*s and one *RGL1*), six putative homologs of ubiquitin E3 ligases (PHOR1s), and several GA-responsive genes including *GL1*, *GAMYB*, *AGAMOUS* and *LUE1*. Moreover, ESTs encoding putative cotton homologs in the auxin biosynthetic (*YUCCA*s, *CYP83B1*s and *NIT2*), signaling (*ARF*s, *AUX1*, *TIR1* and *PIN1*) and transport (*AUX1* and *PIN1*) pathways (Weijers and Jurgens, 2005; Woodward and Bartel, 2005) were enriched in the GH_TMO library (Figure 5b and Table S3).

In addition to the enrichment of positive phytohormonal regulators, putative negative modulator genes such as *IAA26*, *PKS3*, *ROP2*, *ROP6*, *ABI1* and *ABI2* in the auxin and ABA response pathways were also enriched in the libraries constructed from young ovules (GH_GH_TMO) or elongating fibers (GH_BNL and GA_Ea; Figure 4b and Table S3).

Quantitative RT-PCR analysis confirmed the transcript abundance of selected ESTs encoding putative phytohormonal regulators in ovules at early stages of fiber development (−3 to +3 DPA, Figure 5e–l). A subset of eight genes encoding putative phytohormonal regulators was upregulated in the immature and fiber-bearing ovules but down-regulated in leaves and petals and in the *N1N1* naked seed mutant. These genes encode a putative BR receptor, BRI1 (*TC77907*), a BR positive regulator, BES1 (*TC65017*), an auxin receptor, TIR1 (*TC69611*), three auxin response factors, ARF1 (*TC76794*), ARF8 (*DT565265*) and ARF12 (*TC64087*), an ethylene response factor (ERF1, *TC66418*), and a negative regulator of GA (GAI, *TC70523*). It is notable that all eight phytohormonal pathway-related genes were expressed at least one stage earlier than three of four MYB-related genes, with the exception of DT553497 (Figure 5). Moreover, the expression of phytohormonal pathway-related genes declined at a later stage (5 DPA).

Enrichment of genome-specific transcripts in the GH_TMO ESTs

The cultivated allotetraploid cotton species, *G. hirsutum* L. and *G. barbadense* L. (AADD), were formed by allopolyploidization between two diploid cotton species, *G. arboreum* L. (or *G. herbaceum* L.; AA) and *G. raimondii* Ulbrich (DD; Beasley, 1940; Wendel and Cronn, 2003). Notably, the AA genome donor produces long lint fibers, whereas the DD genome donor produces a very few short fibers, suggesting a role for genome-specific gene expression in fiber cell development.

To determine transcript origin in *G. hirsutum* L., we analyzed TCs in CGI7 with at least five ESTs for the presence of GSPs. A total of 32 229 GSPs were identified (Table S4);

Figure 4. Hierarchical cluster analyses for cotton ESTs encoding (a) putative transcription factors and (b) phytohormonal pathway-related proteins that were determined to be differentially expressed in five cotton EST libraries derived from tissues at different developmental stages using the method described by Stekel *et al.* (2000) and a *P* value of 0.001.
The frequencies of ESTs from each library in each TC are represented by increasing intensities of red (black represents a frequency of zero). Gene identities in the parentheses correspond to (a) transcription factor gene families in Arabidopsis and (b) homologous phytohormonal regulators in Arabidopsis and other species. The libraries were created from GH_TMO (*G. hirsutum* TM-1 ovule), GH_BNL (*G. hirsutum* six-day cotton fibers), GA_Ea (*G. arboreum* developing fibers, 7–10 DPA), GR_Ea (*G. raimondii* whole seedlings with the first true leaves), and GR_Eb (*G. raimondii* flower buds and bolls; buds −3 DPA to +3 DPA).

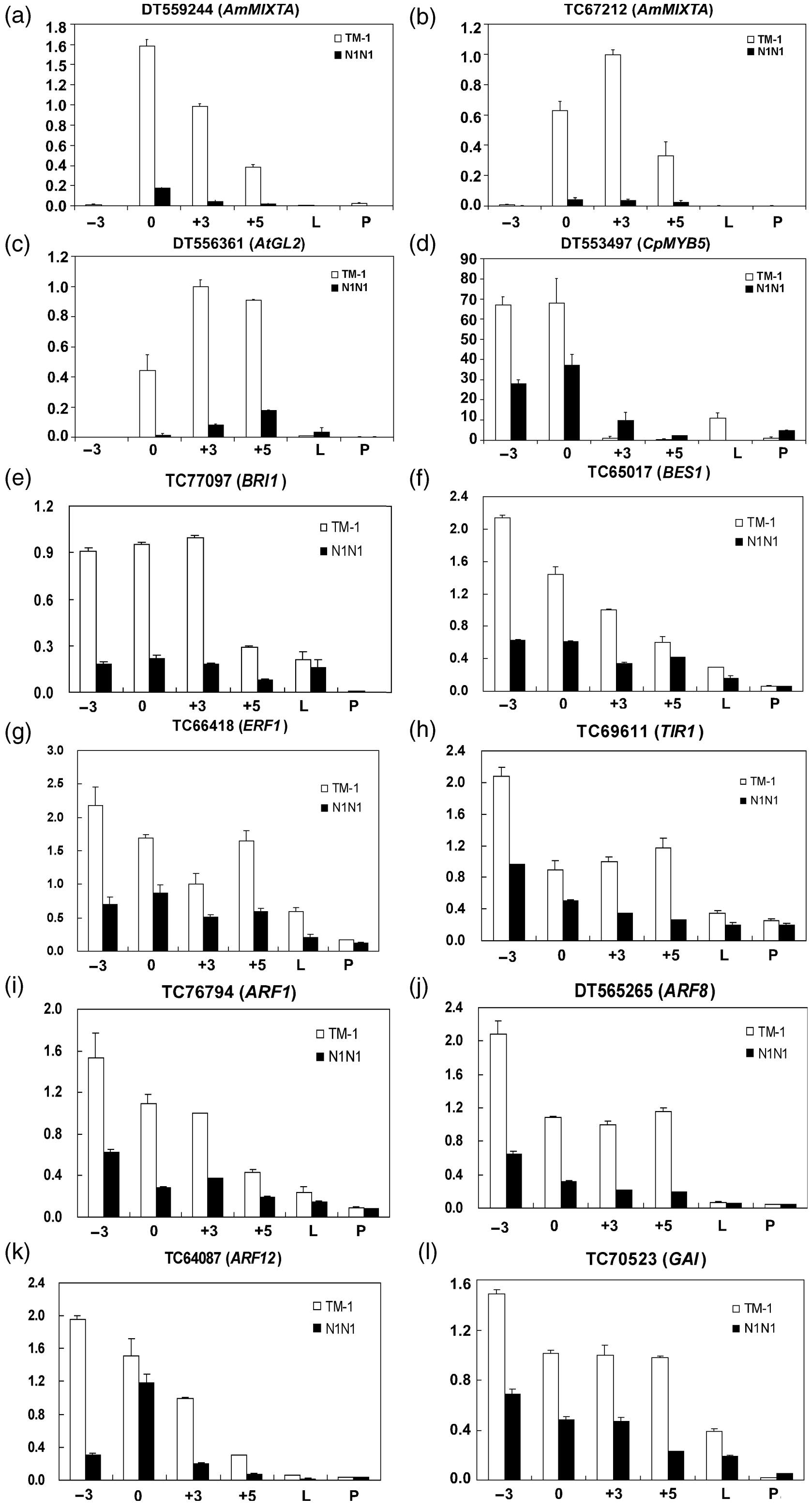


Figure 5. Quantitative RT-PCR analyses of selected genes encoding putative GH_TMO-specific transcription factors and phytohormonal regulators.

(a) *DT559244*, a putative *AmMIXTA* homolog.

(b) *TC67212*, a putative *AmMIXTA* homolog.

(c) *DT556361*, an *AtGL2* homolog.

(d) *DT553497*, a putative *CpMYB5* homolog.

(e) *TC77907*, a brassinosteroid (BR) receptor 1 (*BRI1*)-like gene.

(f) *TC65017*, a BR signaling positive regulator-like gene (*BES1*).

(g) *TC66418*, an ethylene response factor 1 (*ERF1*)-like gene.

(h) *TC69611*, an auxin signaling-related gene (*TIR1*).

(i) *TC76794*, an auxin-response factor 1 (*ARF1*)-like gene.

(j) *DT565265*, an *ARF8*-like gene.

(k) *TC64087*, an *ARF12*-like gene.

(l) *TC70523*, a GA negative regulator (GAI)-related gene. TM-1, Texas Marker-1 (wild-type); N1N1, *N1N1* fiberless mutant; −3, 3 days before anthesis; 0, the day of anthesis; +3, 3 days post-anthesis (DPA); +5, 5 DPA; L, leaves; P, petals.

approximately 64% of the GSPs were transitions (A/G and T/C) and approximately 36% were transversions (G/T, A/C, C/G and A/T). Among the GSPs identified, 10 066 (approximately 31%) were consistently linked (100% out of at least 5 ESTs examined) to the AA or DD genotype in all AA, DD and AADD ESTs within each TC (Figure 6a and Table S5). The numbers of ESTs derived from the AA or DD genomes were equally distributed in the CGI7 assembly. In contrast, many tran-

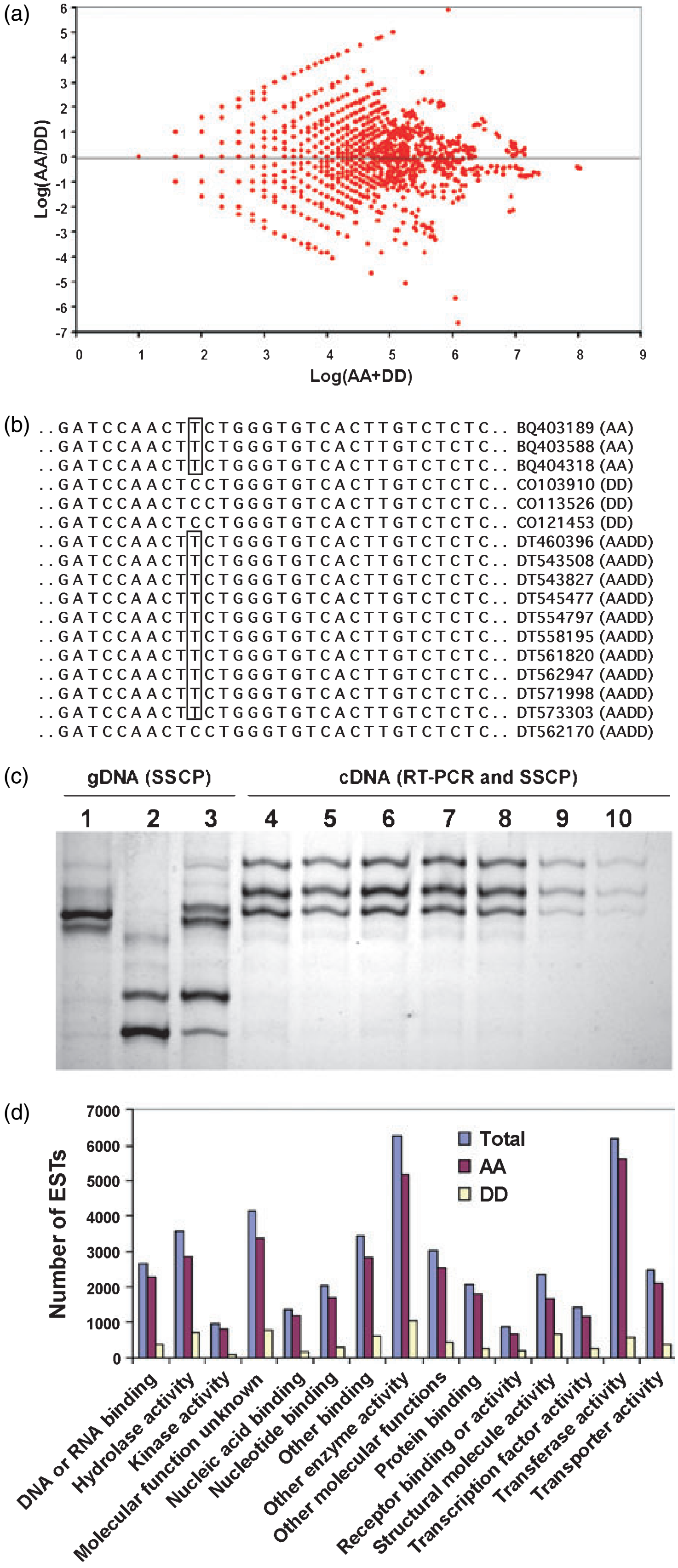


Figure 6. Genome-specific gene regulation in cotton allotetraploids.

(a) A scatter plot representing the AA versus DD transcript ratio among allotetraploid cotton ESTs that contained AA- or DD-specific GSPs. For each TC, the $\log_2$(AA/DD) value was plotted against the $\log_2$(AA+DD) value. The AADD transcripts with AA or DD origins that were preferentially accumulated are located above or below, respectively, the zero horizontal line. AA/DD, the number of AA ESTs divided by the number of DD ESTs in AADD cotton in each TC; AA+DD, the sum of AA ESTs and DD ESTs in AADD cotton in each TC.

(b) An example of AA and DD-EST (*TC67135* encoding putative cotton cyclin D) distributions in *G. arboreum* (AA), *G. ranmondii* (DD) and *G. hirsutum* (AADD). A T/C GSP was detected between AA and DD ESTs. An equal number of ESTs were present in AA and DD genotypes, but nine of ten ESTs in the AADD genotype were AA-specific (boxed).

(c) SSCP analysis confirmed genotype-specific expression patterns of *TC67135* (*cyclin D*). Lanes 1–10 indicate PCR products amplified from genomic DNA of AA (lane 1), DD (lane 2) and AADD (lane 3), and cDNAs from ovules harvested at −3 DPA (lane 4), 0 DPA (lane 5), +3 DPA (lane 6), +5 DPA (lane 7) and +7 DPA (lane 8), leaves (lane 9) and petals (lane 10), all from that AADD genotype. Note that the transcripts were AA-specific and expressed at high levels in immature and fiber-bearing ovules.

(d) GO functional classifications of ESTs derived from AA and DD sub-genomes. The frequency of AA sub-genome ESTs (purple) is significantly higher than that of DD sub-genome ESTs (yellow). The total number of ESTs is shown by the blue bars.

scripts present in TM-1 allotetraploid cotton were derived from only one diploid genotype (Figure 6a and Table S6). Among 2605 TCs examined, 2233 were derived exclusively from the AA sub-genome and 372 from the DD sub-genome. For example, *TC67135*, a putative cotton cyclin D3 gene, contained three ESTs with a T and three with a C at position 638 in AA and DD genome cotton, respectively (Figure 6b). Allotetraploid (AADD) cotton contained 10 ESTs with the T and one with the C at the same position, suggesting preferential expression of this gene from the AA sub-genome in TM-1. TC66418, a putative cotton *ERF1*, a gene involved in ethylene signaling, had two ESTs with an A and 16 with a G at position 376 in AA and DD genome cotton, respectively. In allotetraploid (AADD) cotton, all 50 ESTs for the putative *ERF1* were derived from the AA sub-genome because there was an A at position 376 (data not shown).

We confirmed the preferential expression patterns of a few genes using SSCP analysis (*TC67135*, Figure 6c). The differentially accumulated transcripts in the AA sub-genome in cotton allopolyploids included the ESTs encoding putative transcription factors such as MYB, WRKY and bZIP families, and regulators involved in phytohormone signal transduction pathways. Furthermore, AA sub-genome transcripts were significantly over-represented in every molecular function class (Figure 6d). Thus, our data documented sequential activation (Lee *et al.*, 2006) of genome-wide AA sub-genome-specific transcriptional and phytohormonal genes involved in biological processes important for early stages of fiber cell development.

Discussion

Polyploidy effects of gene regulation in cotton allotetraploids

Results from ovular EST analysis indicate preferential accumulation of AA sub-genome ESTs encoding putative transcription factors and phytohormonal regulators in the cotton allotetraploids containing AADD genomes. The AA genome species produce both fuzz and long lint fibers, whereas the DD genome species produce very short lint fibers (Applequist *et al.*, 2001), although the DD genome contributes to other agronomic traits. The data suggest that the expression of genes from the fiber-bearing AA genome donor contributes to early stages of fiber and seed development. Alternatively, the DD sub-genome may contain *trans*-activating factors that stimulate the expression of genes in the AA sub-genome. Using computational, RT-PCR and SSCP analyses, we identified and verified progenitor-dependent (or genome-specific) gene regulation in allotetraploid cotton (Figure 6), which is reminiscent of the genome-wide transcriptome dominance (*A. arenosa* over *A. thaliana*) discovered in Arabidopsis allotetraploids (Wang *et al.*, 2006). Therefore, genome-specific gene regulation may be a general consequence of polyploidization in Arabidopsis and cotton allopolyploids and presumably in many other allopolyploid plants.

Duplicate genes in cotton allopolyploids may diverge their expression via tissue-specific regulation or sub-functionalization (Adams *et al.*, 2003, 2004). Among 40 genes examined in *G. hirsutum* ovules, 10 displayed silencing or unequal expression of homoeologous loci. Although silencing is reciprocal and developmentally regulated, there is no bias towards either the AA or DD sub-genome. This is probably because many of the genes studied were randomly selected and examined in non-fiber tissues. It is likely that genome-dependent transcript accumulation is specific to a particular developmental stage (e.g. fiber cell initiation). Indeed, the genes encoding putative MYB transcription factors and phytohormonal regulators in ovules were expressed at high levels in the early stages but their expression levels declined in later stages of fiber development (Figure 5). Accumulation of AA genome-specific transcripts in the allotetraploid cotton may facilitate the fiber-specific gene regulation that has been selected after polyploidization and during domestication of cotton allotetraploids. The data also suggest a mechanism for how cultivated cotton has a fiber cell phenotype that is similar to the AA genome progenitor. The low expression levels of the genes encoding putative transcription factors and phytohormonal regulators in the *N1N1* mutant (Figure 5) suggest that these genes contribute to the development of the long lint fibers that are significantly reduced in number in the *N1N1* mutant.

A previous study indicated that the majority of fiber quantative trait loci (QTLs) are located in the DD sub-genome (Jiang *et al.*, 1998), probably because some genes associated with fiber development are suppressed in the DD genome diploid but de-repressed after combining the AA and DD sub-genomes in the allotetraploids. The combination of AA and DD homoeologous genomes stimulates production of superior fibers compared with the AA genome donor. Fiberless mutations in the *G. hirsutum* allotetraploid are thought to be located in similar positions relative to the centromeres of AA and DD homoeologous chromosomes in the allotetraploid (Samora *et al.*, 1994), suggesting that novel gene expression is activated after combining homoeologous chromosomes. The expression of genes in the AA sub-genome is enhanced because of interactions between the homoeologous chromosomes, a finding in agreement with the many QTLs identified in the AA sub-genome (Frelichowski *et al.*, 2006; Lacape *et al.*, 2005; Mei *et al.*, 2004; Ulloa *et al.*, 2005; Zhang *et al.*, 2003).

Upregulation of transcription factors during early stages of fiber and ovule development

Compared with other libraries, GH_TMO ESTs represent >15% of the TCs and singletons in CGI7 and contain approximately 4.8% fiber- and ovule-specific transcripts. Note that the GH_TMO ESTs contain different profiles of ESTs encoding putative transcription factors and phytohormonal regulators than those derived from ovules at 5–10 DPA (Shi *et al.*, 2006). The genes encoding these ESTs may play important roles in the differentiation of fiber cell initials in the protodermal layer of immature ovules. For example, *GhPDF1*, a gene encoding a putative protodermal factor 1 in cotton, is highly expressed in immature ovules (–3 DPA) and in fiber-bearing ovules (+5 DPA; Table 1; Lee *et al.*, 2006). In Arabidopsis, *PDF1* is specifically expressed in the L1 layer of vegetative and floral meristems, in organ primordia, and in protodermal cells during embryogenesis, suggesting a role in cell fate determination (Abe *et al.*, 2001).

We found that almost every category of transcription factor is over-represented in the GH_TMO ESTs that were generated from tissues undergoing rapid cellular and developmental transitions during early seed development and fiber cell initiation. Among them, *GL1*, an R2R3-MYB transcription factor, plays a role in leaf trichome differentiation in Arabidopsis (Glover, 2000; Hülskamp, 2004; Ramsay and Glover, 2005). *AtGL2* (At1g79840) is a homeodomain leucine zipper protein that activates downstream trichome-specific differentiation genes during leaf trichome development in Arabidopsis (Rerie *et al.*, 1994). The role of cotton MYB transcription factors in cotton fiber development has also been documented (Loguercio *et al.*, 1999; Suo *et al.*, 2003; Wang *et al.*, 2004). Overexpressing *GaMYB2* complements the *gl1* mutant phenotype and promotes the development of a single seed trichome in Arabidopsis (Wang *et al.*, 2004). *GhMYB25* is expressed in fiber cell initials (Wu *et al.*, 2006).

We confirmed that four genes encoding two putative MYB transcription factors (AmMIXTA and CpMYB5) and a putative homeodomain leucine zipper protein (GL2) were expressed at high levels during early stages (–3 to +5 DPA) of fiber development, and all were repressed in the *N1N1* mutant. *AtGL2* is involved in trichome cell patterning in Arabidopsis leaves (Glover, 2000; Hülskamp, 2004; Ramsay and Glover, 2005), and AmMIXTAs control trichome differentiation in *Antirrhinum* petals (Noda *et al.*, 1994; Perez-Rodriguez *et al.*, 2005). Notably, several putative cotton *MYB* genes form their own clades (Figure 3), many of which are present only in the GH_TMO library (Figure 4a). Enrichment of these putative MYB factors supports their roles in fiber cell differentiation and seed development.

TTG2/AtWRKY44 (*TRANSPARENT TESTA GLABRA2*), the first WRKY gene to be functionally characterized, controls trichome development in leaves and the production of mucilage and tannin in Arabidopsis seed coats (Johnson *et al.*, 2002). The trichomes in the *ttg2* mutants are unbranched and reduced in number compared with the wild-type. WRKY transcription factors were the most abundant transcription factor family in CGI7. The WRKY families participate in various hormonal signaling pathways (Ülker and Somssich, 2004). Some WRKYs are positive regulators of the abscisic acid (ABA) signaling pathway, whereas others such as those in rice aleurone cells are negative regulators of the GA pathway (Xie *et al.*, 2005; Zhang *et al.*, 2004). The data may suggest cross-interactions between transcriptional and phytohormonal regulation during fiber cell development.

The role of phytohormone regulators in early stages of fiber development

Fibers can be induced from the protodermal cells of unfertilized ovules by the addition of auxin and gibberellic acid (GA; Beasley and Ting, 1974; Gialvalis and Seagull, 2001; Kim and Triplett, 2001). Either hormone promotes fiber initiation, and the effect is additive, implying a significant role for the two phytohormones in fiber cell initiation. Moreover, exogenous application of IAA and GA_3 to flower buds *in planta* and unfertilized ovules *in vitro* resulted in an increased number of fibers (Gialvalis and Seagull, 2001). The brassinosteroid brassinolide (BL) promotes fiber cell elongation as well as fiber cell initiation (Sun *et al.*, 2005).

The GH_TMO library contained seven putative DELLA-like genes (four *RGL2*s, two *GAI*s and one *RGL1*), six putative homologs of ubiquitin E3 ligases (PHOR1s), and several GA-responsive genes including *GL1*, *GAMYB*, *AGAMOUS* and *LUE1*. ESTs encoding putative cotton homologs in the auxin biosynthetic (*YUCCA*s, *CYP83B1*s, *SIR1* and *NIT2*), signaling (*ARF*s, *AUX1*, *TIR1* and *PIN1*) and transport (*AUX1* and *PIN1*) pathways (Weijers and Jurgens, 2005; Woodward and Bartel, 2005) were enriched in the GH_TMO library (Figure 5b and Table S3).

Quantitative RT-PCR results confirmed upregulation of a DELLA-like gene (*GAI*, *TC70523*), an auxin receptor (*TC69611*) and three auxin responsive-like factors (*TC64087*, *TC76794* and *DT565265*) in the immature and fiber-bearing ovules (–3 to +5 DPA) compared with non-fiber tissues such as leaves and petals. These genes were significantly downregulated in the *N1N1* naked seed mutant that lacks long lint fibers (Figure 5). Members of the DELLA family repress the GA signaling pathway in the absence of GA and are rapidly degraded by the ubiquitin-26S-proteasome pathway (Fleet and Sun, 2005). Similar to GA, the auxin signal transduction pathway is activated through a de-repression mechanism (Berleth *et al.*, 2004; Weijers and Jurgens, 2005; Woodward and Bartel, 2005). Auxin plays a pivotal role in the establishment of embryonic cell fate along the apical–basal axis during Arabidopsis embryogenesis by localized rearrangement of auxin transporters and responsive factors (Kepinski and Leyser, 2002). Furthermore, cotton genes encoding the putative brassinosteroid (BR) receptor 1 (BRI1), BR signaling positive regulator (BES1) and ethylene response factor 1 (ERF1) were upregulated in the fiber-bearing ovules (Figure 5), suggesting a role for BR signaling and ethylene response genes in fiber cell development (Shi *et al.*, 2006; Sun *et al.*, 2005). Interestingly, all eight phytohormonal pathway-related genes examined were induced prior to the activation of *MYB*-related genes (Figure 5), indicating that phytohormonal pathways potentiate transcriptional regulation for cell fate determination. The spatial and temporal regulation of gene expression during fiber cell development is genome-specific, which is consistent with the long lint fibers produced in AA genome species. The GH_TMO ESTs will be valuable for transcriptome analysis of these genes in immature ovules and young fibers produced *in vitro* and *in vivo*. This will lead to an understanding of how AA diploid progenitors maintain and promote genome-specific gene expression patterns, and how gene regulation between homoeologous AA and DD sub-genomes contributes to enhanced fiber morphology in the cultivated cotton allopolyploids.

Experimental procedures

A full-length cDNA library of Gossypium hirsutum L. Texas Marker-1 (TM-1) ovules (GH_TMO)

We grew *G. hirsutum* L. Texas Marker-1 (TM-1) in a greenhouse at Texas A&M University and in the field at The US Department of Agriculture-Agriculture Research Service (USDA-ARS) (New Orleans, LA, USA). Ovules were harvested at 3 days before anthesis (−3 DPA), the day of anthesis (0 DPA), and 3 days post-anthesis (+3 DPA) from a pool of ~100 plants. Total RNA was extracted from immature ovules using a published method (Chang *et al.*, 1993). An equal amount of RNA was pooled from three samples for each stage and sent to Invitrogen Corp. (Carlsbad, CA, USA) for full-length cDNA library construction. Briefly, mRNA was isolated from the total RNA using a filter syringe containing oligo (dT). The first-strand cDNA was synthesized from 15 μg mRNA using SuperScript^{TM} III reverse transcriptase, and the second-strand cDNA was synthesized using *Escherichia coli* RNase H, DNA polymerase I and DNA ligase. The double-stranded cDNA was blunt-ended using T4 DNA polymerase, digested with *Not*I, size-selected using agarose gel electrophoresis, and directionally cloned into the *Not*I–*Eco*RV sites of pCMV·SPORT6.1. The ligated products were transformed using *ELECTROMAX*^{TM} DH10BT1 cells. After the library was plated, 23 clones were randomly selected for quality control. The colonies were arrayed in 384-well plates (51 072 clones in 133 384-well plates) in duplicate sets. One set was sent to the Institute for Genomic Research (Rockville, MD, USA) for sequencing, and the other set was stored in a −70°C freezer.

EST sequencing and data analysis

Approximately 40 000 cDNA clones were single-pass sequenced primarily from the 5′ end at the Institute for Genomic Research (http://www.tigr.org) using the SP6 primer. Plasmid DNA was prepared using a modified alkaline lysis procedure, and sequenced using Big Dye^{TM} Terminator Cycle Sequencing Ready Reaction reagents and ABI 3730xl Genetic Analyzers (Applied Biosystems, Foster City, CA, USA). Following electrophoresis and fluorescence detection, quality values were assigned to each base call by TraceTuner software (Paracel Inc., Pasadena, CA, USA), and LUCY (Chou and Holmes, 2001) was used to automatically process the raw sequence data, including trimming of low-quality bases, vector and *E. coli* sequences, and those sequences <100 bp. The procedures for EST cleaning, assembly into tentative consensus (TC) groups, and annotation have been described previously (Quackenbush, 2001; http://www.tigr.org/tdb/tgi/gifaq.shtml). ESTs were grouped into clusters by sequence similarity and clone links using the TIGR Gene Index Clustering tools (TGICL) clustering utilities (Pertea *et al.*, 2003). Each cluster was assembled at high stringency using the Paracel Transcript Assembler (Paracel Inc.) to produce TCs. Those ESTs that did not assemble into a TC were termed 'singletons'.

The gene ontology (GO) molecular functional class for each cotton sequence generated was assigned based on Arabidopsis GO SLIM (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.20050723.txt). The cotton unique sequences were compared with the Arabidopsis proteome using BLASTX with an *E*-value ≤ −10. The GO functional class for the top Arabidopsis hit was assigned to each cotton sequence.

In silico analysis of gene expression in five different cotton EST libraries was performed using the method described by Stekel *et al.* (2000). An expression profile matrix was built representing the frequency of EST within each TC in each library, and imported into the IDEG6 software (Romualdi *et al.*, 2001, 2003) to calculate *R* statistics (Stekel *et al.*, 2000). Differentially expressed TCs were identified using a *P* value of 0.001, and those with at least five ESTs were imported into MultiExperiment Viewer (MEV, The Institute for Genomic Research, www.tigr.org/software/microarray.shtml) for hierarchical cluster analysis (Eisen *et al.*, 1998).

To identify the presence of putative cotton candidate genes, the amino acid sequences of Arabidopsis genes involved in various biological pathways were downloaded from GenBank and compared with cotton EST sequences using TBLASTN with an *E*-value ≤ −10. Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 3.0 (Kumar *et al.*, 2004).

Analysis of genome-specific polymorphisms (GSPs) and simple sequence repeats (SSRs)

The most extensively cultivated cotton species, *G. hirsutum* and *G. barbadense*, are allotetraploids derived by combining two genomes probably from *G. arboreum* L. and *G. ramondii* L. ancestral species. Interestingly, if both genomes are transcriptionally active in the allopolyploid, the GH_TMO library would be expected to contain ESTs derived from two homoeologous genomes. To estimate the relative abundance of the A- and D-genome ESTs in the library, we examined the presence of genome-specific polymorphisms (GSPs) among TCs with at least five ESTs in the Cotton Gene Index version 7 (CGI7, www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=cotton). To avoid scoring sequencing errors as GSPs, we required that polymorphic nucleotides be present in at least two different libraries. AA- or DD-specific GSPs that could be consistently (100%) linked to the AA or DD genotype were identified. The criteria for classifying AA or DD GSPs were as follows: (1) more than one AA EST carrying one GSP, (2) more than one DD EST carrying the alternative GSP, (3) more than one AADD EST carrying either the AA or DD GSP, and (4) each GSP from the AADD genotype must be 100% consistent with the AA or DD GSP. Using these criteria, we inferred the origins of ESTs from either the AA or DD genome in the GH_TMO library.

CGI7 sequences were analyzed for the presence of Simple Sequence Repeats (SSRs) using MISA (Thiel *et al.*, 2003). The minimum numbers of repeats specified were 20 for mononucleotides, 10 for dinucleotides, seven for trinucleotides, and five for tetra- to hexanucleotides. The maximum length of interruption allowed between two repeats for compound repeats was 100 bp. Primers spanning each SSR were also designed automatically by analyzing the MISA output using Primer3 (http://www-genome.wi.mit.edu/genome_software/) using Perl script provided by MISA (http://pgrc.ipk-gatersleben.de/misa/primer3.html). The parameters for SSR detection and primer design were as previously described (Kuhl *et al*., 2004).

A total of 3977 SSRs were identified among 2931 sequences, representing approximately 5.3% of the total unique sequences (55 673) in CGI7. The estimated frequency of SSRs among expressed sequences was one SSR per 11 kb, more frequent than the frequency of one SSR per 20 kb observed in a previous report (Cardle *et al*., 2000). Mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats comprised about 70.4%, 9.6%, 13.3%, 3.8%, 1% and 1.9% of the SSRs in CGI7, respectively.

Quantitative RT-PCR

Total RNAs were subjected to treatment with RNase-free DNase I (Ambion Inc., Austin, TX, USA) to remove residual DNA. The first-strand cDNAs for each sample were generated using random hexamers and Taqman reverse transcription reagents (Applied Biosystems). Gene-specific primers were designed based on sequences from the CGI7 database using Primer Express version 2.0 (Applied Biosystems). Samples and standards were run in duplicate on each plate. The quantitative PCR was repeated on at least two plates using an ABI7500 Sequence Detection System (Applied Biosystems). Real-time RT-PCR was performed in a 20 μl reaction mixture containing 7 μl ddH_2O, 10 μl 2x PCR mix, 1 μl forward primer (1 μM), 1 μl reverse primer (1 μM) and 1 μl of template cDNA (10 ng μl^{-1}). The PCR conditions were 2 min pre-incubation at 50°C, 10 min pre-denaturation at 94°C, 40 cycles of 15 sec at 95°C and 1 min at 60°C, followed by steps for dissociation curve generation (30 sec at 95°C, 60 sec at 60°C and 30 sec at 95°C). The 7500 System SDS software version 1.2.2. (Applied Biosystems) was used for data collection and analysis. Dissociation curves for each amplicon were carefully examined to confirm the specificity of the primer pair used. Relative transcript levels for each sample were obtained using the comparative C_T method (Livak and Schmittgen, 2001). The threshold cycle (C_T) value obtained after each reaction was normalized to the C_T value of 18S rRNA. The relative expression level was obtained by calibrating the $\Delta\Delta C_T$ values for other samples using a normalized C_T value ($\Delta\Delta C_T$) for TM-1 (+3 DPA).

The primer pairs are 18S-forward AAGACGGACCACTGCGAAAG and 18S-reverse ATCCCTGGTCGGCATCGT; TC76794-forward GCCTCCTGCGCCTCAA and TC76794-reverse GAGGCAGTGAGGGTCTTGCA; DT565265-forward GCATAACACGACTCAAAGTGATCAG and DT565265-reverse TTTCAGAAATGATGAAGGCACATT; TC69611-forward ACCCTTGGCTGGAAAAAGTTC and TC69611-reverse GGCAAGTGTTGCCAGATCATC; TC65017-forward GGCTTGAGGCAATACGGTAATT and TC65017-reverse AGGCGGCTAGCACATCGTT; TC77907-forward CTTAAGACGCTTGATCTCAGCTACA and TC77907-reverse CGGCAAGGTTCCGATTGA; TC70523-forward GCGGCTTCCAGGCTTGA and TC70523-reverse ACTGCCACCGATTCTTTTGG; TC66418-forward CCAAGTTCCGACAATGTTATGC and TC66418-reverse CCAACAGCTCGTCCGACAA; DT553497-forward TTGCACAACACTTGCCTGGA and DT553497-reverse GCTGTTTCGCCTGTTTTTGG; DT559244-forward CAACCGGTAACCCCAAGGAT and DT559244-reverse GCCTGGCTTC GGCTTCTAAC; TC67212-forward CCTAAAACCGATGCACTCGG and TC67212-reverse ACTCTCCCATTGAGCCATGTG; DT556361-forward AGAGACGTTATGGATTGCGGA and DT556361-reverse TCGTGAAAGGTCGGAGGAAT; TC64087-forward CGATGTTGGCAAGCTGAACA and TC64087-reverse GTCACTGCCTAGCGGGAAGT; and TC67135-forward TGTATCAACAAGACCCCAGGC and TC67135-reverse CGCATTACTTGCTCTTGGGC (for SSCP analysis, see below).

Single-strand conformation polymorphism (SSCP) analysis

SSCP gel electrophoresis was performed in a 12% polyacrylamide gel, with 1x TBE, 10% ammonium persulfate and 0.05% v/v TEMED. Each PCR product (approximately 200 bp) was pre-examined in a 1% agarose gel, mixed with an equal volume of 2x SSCP gel loading dye (0.05% bromophenol blue, 0.05% xylene cyanol, 95% formamide, 20 mM EDTA), denatured at 94°C for 5 min, and cooled on ice prior to loading. Electrophoresis was performed in 0.5x TBE at 200 V for approximately 16 h at room temperature. The gel was fixed in 10% acetic acid for 30 min, washed with deionized H_2O three times for 5 min each, incubated in silver staining solution (0.1% silver nitrate and 0.15% formaldehyde) for 30 min, briefly washed in H_2O for 5 sec, and incubated in pre-cooled developing solution (3% Na_2CO_3, 0.15% formaldehyde, 0.024% $Na_2S_2O_3$). When clear bands appeared, the developing solution was decanted, and 10% acetic acid was added to stop development. Gel images were taken using a CCD camera Gel Logic 100 (Kodak, Rochester, NY, USA).

Acknowledgements

We thank Andrew Woodward for critical suggestions on improving the manuscript. This work is supported by a grant from the National Science Foundation Plant Genome Research Program (DBI0624077). Work in the Chen laboratory is supported in part by Cotton Incorporated (04-5555) and the Texas Higher Education Coordinating Board Advanced Technology Program (000517-0218-2001).

Supplementary Material

The following supplementary material is available for this article online:
Table S1 The ESTs enriched in five-cotton CDMA libraries
Table S2 Putative transcription factors genes in GH_TMO
Table S3 Phytohormonal related genes in GH_TMO library
Table S4 SSRs and primers in GGI7
Table S5 GSPs in AA, DD and AADD ESTs
Table S6 AA or DD GSPs in AADD ESTs
This material is available as part of the online article from http://www.blackwell-synergy.com

References

Abe, M., Takahashi, T. and Komeda, Y. (2001) Identification of a cis-regulatory element for L1 layer-specific gene expression, which is targeted by an L1-specific homeodomain protein. *Plant J.* **26**, 487–494.

Abo-el-Saad, M. and Wu, R. (1995) A rice membrane calcium-dependent protein kinase is induced by gibberellin. *Plant Physiol.* **108**, 787–793.

Adams, K.L., Cronn, R., Percifield, R. and Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl Acad. Sci. USA*, **100**, 4649–4654.

Adams, K.L., Percifield, R. and Wendel, J.F. (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics*, **168**, 2217–2226.

Applequist, W.L., Cronn, R. and Wendel, J.F. (2001) Comparative development of fiber in wild and cultivated cotton. *Evol. Dev.* **3**, 3–17.

Arpat, A.B., Waugh, M., Sullivan, J.P., Gonzales, M., Frisch, D., Main, D., Wood, T., Leslie, A., Wing, R.A. and Wilkins, T.A. (2004) Functional genomics of cell elongation in developing cotton fibers. *Plant Mol. Biol.* **54**, 911–929.

Basra, A. and Malik, C.P. (1984) Development of the cotton fiber. *Int. Rev. Cytol.* **89**, 65–113.

Beasley, J.O. (1940) The origin of American tetraploid *Gossypium* species. *Am. Nat.*, **74**, 285–286.

Beasley, C.A. and Ting, I.P. (1974) The effects of plant growth substances on *in vitro* fiber development from unfertilized cotton ovules. *Am. J. Bot.* **61**, 188–194.

Berleth, T., Krogan, N.T. and Scarpella, E. (2004) Auxin signals – turning genes on and turning cells around. *Curr. Opin. Plant Biol.* **7**, 553–563.

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. and Waugh, R. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics*, **156**, 847–854.

Chang, S., Puryear, J. and Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **11**, 113–116.

Chou, H.H. and Holmes, M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**, 1093–1104.

Delmer, D.P., Pear, J., Andrawis, A. and Stalker, D. (1995) Genes encoding small GTP-binding proteins analogous to mammalian RAC are preferentially expressed in developing cotton fibers. *Mol. Gen. Genet.* **248**, 43–51.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Endrizzi, J.E., Turcotte, E.L. and Kohel, R.J. (1984) Qualitative genetics, cytology and cytogenetics. In *Agronomy: Cotton* (Kohel, R.J. and Lewis, D.F., eds). Madison, WI: American Society of Agronomy, Inc., pp. 59–80.

Fleet, C.M. and Sun, T.P. (2005) A DELLAcate balance: the role of gibberellin in plant morphogenesis. *Curr. Opin. Plant Biol.* **8**, 77–85.

Frelichowski, J.E., Jr, Palmer, M.B., Main, D., Tomkins, J.P., Cantrell, R.G., Stelly, D.M., Yu, J., Kohel, R.J. and Ulloa, M. (2006) Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. *Mol. Genet. Genomics*, **275**, 479–491.

Gialvalis, S. and Seagull, R.W. (2001) Plant hormones alter fiber initiation in unfertilized cultured ovules of *Gossypium hirsutum*. *J. Cotton Sci.* **5**, 252–258.

Glover, B.J. (2000) Differentiation in plant epidermal cells. *J. Exp. Bot.* **51**, 497–505.

Haigler, C.H., Zhang, D.H. and Wilkerson, C.G. (2005) Biotechnological improvement of cotton fibre maturity. *Physiol. Plant.* **124**, 285–294.

Hedden, P. and Kamiya, Y. (1997) GIBBERELLIN BIOSYNTHESIS: enzymes, genes and their regulation. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 431–460.

Hülskamp, M. (2004) Plant trichomes: a model for cell differentiation. *Nat. Rev. Mol. Cell Biol.* **5**, 471–480.

Humphries, J.A., Walker, A.R., Timmis, J.N. and Orford, S.J. (2005) Two WD-repeat genes from cotton are functional homologues of the *Arabidopsis thaliana* TRANSPARENT TESTA GLABRA1 (TTG1) gene. *Plant Mol. Biol.* **57**, 67–81.

Ji, S.J., Lu, Y.C., Feng, J.X., Wei, G., Li, J., Shi, Y.H., Fu, Q., Liu, D., Luo, J.C. and Zhu, Y.X. (2003) Isolation and analyses of genes preferentially expressed during early cotton fiber development by subtractive PCR and cDNA array. *Nucleic Acids Res.* **31**, 2534–2543.

Jiang, C., Wright, R.J., El-Zik, K.M. and Paterson, A.H. (1998) Polyploid formation created unique avenues for response to selection in *Gossypium. Proc. Natl Acad. Sci. USA*, **95**, 4419–4424.

John, M.E. and Crow, L.J. (1992) Gene expression in cotton (*Gossypium hirsutum* L.) fiber: cloning of the mRNAs. *Proc. Natl Acad. Sci. USA*, **89**, 5769–5773.

John, M.E. and Keller, G. (1995) Characterization of mRNA for a proline-rich protein of cotton fiber. *Plant Physiol.* **108**, 669–676.

Johnson, C.S., Kolevski, B. and Smyth, D.R. (2002) TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *Plant Cell*, **14**, 1359–1375.

Kepinski, S. and Leyser, O. (2002) Ubiquitination and auxin signaling: a degrading story. *Plant Cell*, **14**, S81–S95.

Kim, H.J. and Triplett, B.A. (2001) Cotton fiber growth in planta and *in vitro*. Models for plant cell elongation and cell wall biogenesis. *Plant Physiol.* **127**, 1361–1366.

Kim, H.J. and Triplett, B.A. (2004) Cotton fiber germin-like protein. I. Molecular cloning and gene expression. *Planta*, **218**, 516–524.

Kranz, H.D., Denekamp, M., Greco, R. ***et al.*** (1998) Towards functional characterisation of the members of the R2R3-MYB gene family from *Arabidopsis thaliana*. *Plant J.* **16**, 263–276.

Kuhl, J.C., Cheung, F., Yuan, Q. ***et al.*** (2004) A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. *Plant Cell*, **16**, 114–125.

Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5**, 150–163.

Lacape, J.M., Nguyen, T.B., Courtois, B., Belot, J.L., Giband, M., Gourlot, J.P., Gawryziak, G., Roques, S. and Hau, B. (2005) QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* × *Gossypium barbadense* backcross generations. *Crop Sci.* **45**, 123–140.

Lee, J.J., Hassan, O.S.S., Gao, W. ***et al.*** (2006) Developmental and gene expression analyses of a cotton naked seed mutant. *Planta*, **223**, 418–432.

Li, C.H., Zhu, Y.Q., Meng, Y.L., Wang, J.W., Xu, K.X., Zhang, T.Z. and Chen, X.Y. (2002) Isolation of genes preferentially expressed in cotton fibers by cDNA filter arrays and RT-PCR. *Plant Sci.* **163**, 1113–1120.

Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time PCR and Z(-Delta Delta C(T)) method. *Methods*, **25**, 402–408.

Loguercio, L.L., Zhang, J.Q. and Wilkins, T.A. (1999) Differential regulation of six novel MYB-domain genes defines two distinct expression patterns in allotetraploid cotton (*Gossypium hirsutum* L.). *Mol. Gen. Genet.* **261**, 660–671.

Ma, D.P., Liu, H.C., Tan, H., Creech, R.G., Jenkins, J.N. and Chang, Y.F. (1997) Cloning and characterization of a cotton lipid transfer protein gene specifically expressed in fiber cells. *Biochim. Biophys. Acta* **1344**, 111–114.

Mei, M., Syed, N.H., Gao, W., Thaxton, P.M., Smith, C.W., Stelly, D.M. and Chen, Z.J. (2004) Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*). *Theor. Appl. Genet.* **108**, 280–291.

Noda, K., Glover, B.J., Linstead, P. and Martin, C. (1994) Flower colour intensity depends on specialized cell shape controlled by a Myb-related transcription factor. *Nature*, **369**, 661–664.

Orford, S.J. and Timmis, J.M. (1997) Abundant mRNAs specific to the developing cotton fibre. *Theor. Appl. Genet.* **94**, 909–918.

Percival, A.E., Wendel, J.F. and Stewart, J.M. (1999) Taxonomy and germplasm resources. In *Cotton: Origin, History, Technology,* *and Production* (Smith, C.W. and Cothren, J.T., eds). New York: John Wiley & Sons, Inc., pp. 33–63.

Perez-Rodriguez, M., Jaffe, F.W., Butelli, E., Glover, B.J. and Martin, C. (2005) Development of three different cell types is associated with the activity of a specific MYB transcription factor in the ventral petal of *Antirrhinum majus* flowers. *Development*, **132**, 359–370.

Pertea, G., Huang, X., Liang, F. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427.

Ramsay, N.A. and Glover, B.J. (2005) MYB–bHLH–WD40 protein complex and the evolution of cellular diversity. *Trends Plant Sci.* **10**, 63–70.

Reinhart, J.A., Petersen, M.W. and John, M.E. (1996) Tissue-specific and developmental regulation of cotton gene fbl2a: demonstration of promoter activity in transgenic plants. *Plant Physiol.* **112**, 1331–1341.

Rerie, W.G., Feldmann, K.A. and Marks, M.D. (1994) The GLABRA2 gene encodes a homeo domain protein required for normal trichome development in Arabidopsis. *Genes Dev.* **8**, 1388–1399.

Romualdi, C., Bortoluzzi, S. and Danieli, G.A. (2001) Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. *Hum. Mol. Genet.* **10**, 2133–2141.

Romualdi, C., Bortoluzzi, S., D'Alessi, F. and Danieli, G.A. (2003) IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. *Physiol. Genomics*, **12**, 159–162.

Rosin, F.M., Hart, J.K., Horner, H.T., Davies, P.J. and Hannapel, D.J. (2003) Overexpression of a knotted-like homeobox gene of potato alters vegetative development by decreasing gibberellin accumulation. *Plant Physiol.* **132**, 106–117.

Samora, P.J., Stelly, D.M. and Kohel, R.J. (1994) Localization and mapping of the Le1 and Gl2 of cotton (*Gossypium hirsutum* L.). *J. Hered.* **85**, 152–157.

Shi, Y.H., Zhu, S.W., Mao, X.Z., Feng, J.X., Qin, Y.M., Zhang, L., Cheng, J., Wei, L.P., Wang, Z.Y. and Zhu, Y.X. (2006) Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell*, **18**, 651–664.

Smart, L.B., Vojdani, F., Maeshima, M. and Wilkins, T.A. (1998) Genes involved in osmoregulation during turgor-driven cell expansion of developing cotton fibers are differentially regulated. *Plant Physiol.* **116**, 1539–1549.

Stekel, D.J., Git, Y. and Falciani, F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res.* **10**, 2055–2061.

Sun, Y., Veerabomma, S., Abdel-Mageed, H.A., Fokar, M., Asami, T., Yoshida, S. and Allen, R.D. (2005) Brassinosteroid regulates fiber development on cultured cotton ovules. *Plant Cell Physiol.* **46**, 1384–1391.

Suo, J., Liang, X., Pu, L., Zhang, Y. and Xue, Y. (2003) Identification of GhMYB109 encoding a R2R3 MYB transcription factor that is expressed specifically in fiber initials and elongating fibers of cotton (*Gossypium hirsutum* L.). *Biochim. Biophys. Acta*, **1630**, 25–34.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **106**, 411–422.

Tiwari, S.C. and Wilkins, T.A. (1995) Cotton (*Gossypium hirsutum*) seed trichomes expand via diffuse growing mechanism. *Can. J. Bot.* **73**, 746–757.

Udall, J.A., Swanson, J.M., Haller, K. *et al.* (2006) A global assembly of cotton ESTs. *Genome Res.* **16**, 441–450.

Ülker, B. and Somssich, I.E. (2004) WRKY transcription factors: from DNA binding towards biological function. *Curr. Opin. Plant Biol.* **7**, 491–498.

Ulloa, M., Saha, S., Jenkins, J.N., Meredith, W.R., Jr, McCarty, J.C., Jr and Stelly, D.M. (2005) Chromosomal assignment of RFLP linkage groups harboring important QTLs on an intraspecific cotton (*Gossypium hirsutum* L.) Joinmap. *J. Hered.* **96**, 132–144.

Wang, S., Wang, J.W., Yu, N., Li, C.H., Luo, B., Gou, J.Y., Wang, L.J. and Chen, X.Y. (2004) Control of plant trichome development by a cotton fiber MYB gene. *Plant Cell*, **16**, 2323–2334.

Wang, J., Tian, L., Lee, H.S. *et al.* (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics*, **172**, 507–517.

Weijers, D. and Jurgens, G. (2005) Auxin and embryo axis formation: the ends in sight? *Curr. Opin. Plant Biol.* **8**, 32–37.

Wendel, J.F. and Cronn, R.C. (2003) Polyploidy and the evolutionary history of cotton. *Adv. Agron.* **78**, 139–186.

Wendel, J.F., Schnabel, A. and Seelanan, T. (1995) An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. *Mol. Phylogenet Evol.* **4**, 298–313.

Whittaker, D.J. and Tripplett, B.A. (1999) Gene-specific changes in alpha-tubulin transcript accumulation in developing cotton fibers. *Plant. Physiol.* **121**, 181–188.

Wilkins, T.a. and Arpat, A.B. (2005) The cotton fiber transcriptome. *Physiol. Plant.* **124**, 295–300.

Wilkins, T.A. and Jernstedt, J.A. (1999) Molecular genetics of developing cotton fibers. In *Cotton Fibers* (Basra, A.M., ed.). New York: Hawthorne Press, pp. 231–267.

Woodward, A.W. and Bartel, B. (2005) Auxin: regulation, action, and interaction. *Ann. Bot.* **95**, 707–735.

Wu, Y., Machado, A.C., White, R.G., Llewellyn, D.J. and Dennis, E.S. (2006) Expression profiling identifies genes expressed early during lint fibre initiation in cotton. *Plant Cell Physiol.* **47**, 107–127.

Xie, Z., Zhang, Z.L., Zou, X., Huang, J., Ruas, P., Thompson, D. and Shen, Q.J. (2005) Annotations and functional analyses of the rice WRKY gene superfamily reveal positive and negative regulators of abscisic acid signaling in aleurone cells. *Plant Physiol.* **137**, 176–189.

Yang, G. and Komatsu, S. (2000) Involvement of calcium-dependent protein kinase in rice (*Oryza sativa* L.) lamina inclination caused by brassinolide. *Plant Cell Physiol.* **41**, 1243–1250.

Yin, Y., Wang, Z.Y., Mora-Garcia, S., Li, J., Yoshida, S., Asami, T. and Chory, J. (2002) BES1 accumulates in the nucleus in response to brassinosteroids to regulate gene expression and promote stem elongation. *Cell*, **109**, 181–191.

Zhang, T., Yuan, Y., Yu, J., Guo, W. and Kohel, R.J. (2003) Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. *Theor. Appl. Genet.* **106**, 262–268.

Zhang, Z.L., Xie, Z., Zou, X., Casaretto, J., Ho, T.H. and Shen, Q.J. (2004) A rice WRKY gene encodes a transcriptional repressor of the gibberellin signaling pathway in aleurone cells. *Plant Physiol.* **134**, 1500–1513.